SAND REPORT

SAND2002-2771
Unlimited Release
Printed September 2002

National Nuclear Security Administration Knowledge Base Contributor's Guide

Dorthe Carr

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

SAND2002-2771
Unlimited Release
Printed September 2002

National Nuclear Security Administration *Knowledge Base Contributor's* Guide

Dorthe Carr
Data Exploitation Department
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1138

Abstract

The National Nuclear Security Administration is creating a *Knowledge Base* to store technical information to support the United States nuclear explosion monitoring mission. This guide is intended to be used by researchers who wish to contribute their work to the *Knowledge Base*. It provides definitions of the kinds of *data sets* or *research products* in the *Knowledge Base*, acceptable data formats, and templates to complete to facilitate the documentation necessary for the *Knowledge Base*.

Acknowledgements

We wish to thank all the past contributors to the *Knowledge Base*. Your contributions have helped NNSA make high quality deliveries of the *Knowledge Base* to the Air Force Technical Applications Center and refine the guidance for *Contributors* who are working to have their data or other *research products* included in future *Knowledge Base* releases.

In addition, we thank the following people who reviewed this document: Leslie Casey, Mark Harris, Chris Young, Lisa Wilkening, David Gallegos, plus various *Scientific Integrators* and *Product Integrators* from four of the Department of Energy (DOE) National Laboratories: Sandia (SNL), Los Alamos (LANL), Lawrence Livermore (LLNL) and Pacific Northwest (PNNL). These organizations are part of the Nuclear Explosion Monitoring Research & Engineering Program (NEM R&E) http://www.nemre.nn.doe.gov sponsored by the National Nuclear Security Administration's Office of Nonproliferation Research & Engineering (NA-22).

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL8500. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy.

Table of Contents

1.0 Introduction

The National Nuclear Security Administration is creating a *Knowledge Base* (see Glossary, section seven) to store technical information to support the United States nuclear explosion monitoring mission. This guide is intended to be used by researchers who wish to contribute their work to the *Knowledge Base*. It explains standards and formats for *Contributors* to use to expedite the inclusion of their data into the *Knowledge Base*. These standards reflect the policies of the ground-based Nuclear Explosion Monitoring Research & Engineering (GNEM R&E) program and its affiliated NNSA and DOE laboratories, Los Alamos National Laboratory, Lawrence Livermore National Laboratory, Pacific Northwest National Laboratory and Sandia National Laboratories.

Including this Introduction, this guide has eight sections. Section two describes the role of the *Product Integrator* with whom the *Contributors* coordinate, and through whom their *research products* are submitted for inclusion into the *Knowledge Base*. The types of information in the *Knowledge Base* are described in section three, along with the acceptable formats for *research products* to be provided to a *Product Integrator*. Section four lists questions the *Product Integrator* will need answered in order to prepare the appropriate documentation, including metadata, needed for any completed *research product*. A summary of this document is in section five, references are in section six, and a small Glossary of Terms is in section seven. The last section is an Appendix - "*Contributor*'s Research Product Summary for *Product Integrator*s". The items to be addressed in the summary are available as a template from the electronic version of this report. If you are reading a hardcopy of this report, you can find an electronic version at http://www.nemre.nn.doe.gov/nemre/kbase_info.hmtl or the electronic version of the proceedings from the 24th Seismic Research Review in CD form or at http://www.nemre.nn.doe.gov/review2002. Comments about this guide should be referred to the author (dbcarr@sandia.gov).

2.0 Product Integrator

The *Product Integrator* is a scientist who has a good overall understanding of the *Knowledge Base*, its structure, content and use. The primary function of the *Product Integrator* is to determine whether a prospective *Contributor's* results and products should be considered for incorporation into the *Knowledge Base*. If such a determination is made, the *Product Integrator's* duty is to ensure that the information is formatted in an acceptable manner and contains sufficient *metadata* and other supporting information to be usable in the *Knowledge Base*.

A *Contributor* can facilitate integration either by providing information in established formats for the *Knowledge Base*, or providing it in such a way that the *Product Integrator* can get the information into the correct format. This guide is intended to be used in consultation with a *Product Integrator* to provide the necessary information to clarify and expedite the process.

The name of the *Product Integrator* for a specific contract can be obtained from the Coordination web page, http://www.nemre.nn.doe.gov/coordination, which lists NNSA sponsored ground-based NEM sponsored R&E monitoring research contracts/awards as well as relevant contracts/awards funded by other sponsors.

3.0 Data Formats

Research products that are considered for inclusion in the *Knowledge Base* may consist of one or more *data sets*. A *data set* is a collection of data, usually all of the same kind. There are five broad categories of *data sets* in the *Knowledge Base*: event data, parametric grid data, contextual data, supporting information and research tools. In this section we will define each broad category, and provide information on what formats should be used by the *Contributor* to give the information to the *Product Integrator.*

3.1 Event Data

Event data covers *research products* that provide information about previous events recorded by a given monitoring technology or mix of technologies. They include information such as when and where the event took place, what caused the event, and often include the raw sensor waveforms for the event. These events may be used as a reference for comparison with a current event. They may also provide well understood ground truth for further processing tasks. The event data is not complete without its associated network/station information. Network/station information describes the sensor stations that recorded the event and ideally, provides long-term average information about the station performance and its background noise environment.

In the *Knowledge Base*, this information is stored in Oracle database tables following the schema posted on the web site http://www.nemre.nn.doe.gov/nemre/kbase_info.html. More information about these tables can be obtained from *Product Integrators*.

Event data can be provided in a number of ways. The preferred way is to provide the data as an export of the Oracle database tables, but comma or space delimited ASCII text files that include all the information in the table can be loaded into the database. If the event data is being supplied in the form of an event bulletin, the following formats can be used.

CSS3.0 from tables, flatfiles or a combination of the two
Harvard CMT catalog format
NORSAR EXP catalogs format
Nordic bulletin format
HYPO bulletin format (includes hypoinverse, hypo71, hypoellipse)
GSE2.0 bulletin format
ISC bulletin format
USGS bulletin formats (HDF, EHDF, EDR)
International Data Center Reviewed Event Bulletin (REB) format

In addition to the database tables, event data may also include flat files of waveform data and flat files for instrument response. The preferred formats for waveform data are:

SAC Binary (Header version 6)
SEED/MINISEED
GSE2.0 types (INT, CM6, CM8)
CSS3.0 types (e1, e2, i2, g2, s3, t4, f4, s4, s4, i4, t8, f8)

The preferred formats for instrument response files are:

IRIS RESP formats
CSS instrument and sensor flat files referencing any kind of response type (fap or pole-zero).

For any other formats, the *Contributor* should contact their *Product Integrator* to see if the *data set* can be accepted as is or needs to be reformatted.

3.2 Parametric Grid Data

Parametric grid data are data associated with particular geographic locations, which can be interpolated onto grids of arbitrary geographic positions. Examples of types of data which could be used to generate parametric grids are travel-time corrections, surface wave group velocities and amplitude corrections. Ultimately for NNSA software, the interpolation is done using kriging based on geostatistical analysis of observed data, and to facilitate this the data and parameters for the kriging are stored in a custom format designed by the NNSA researchers.

Contributors do not need to use or conform to the custom format, however; they need only ensure that their observation data (travel times, amplitudes, etc.) are delivered to the *Product Integrator* in an ASCII electronic file. Each observation should include a measured value, an error associated with that measurement, and a tie to a ground truth location. The ground truth location must also be provided, and should include hypocentral parameters as well as error estimates for those parameters. The *Product Integrator* will use the NNSA software to perform the geostatistical analysis of the *Contributor's* data and to create the proper format files for use as part of the *Knowledge Base*.

3.3 Contextual Data

Contextual data is a "reference bookshelf" of information intended to provide a context in which to research and evaluate events. It is the broadest ranging of the categories of information providing geographic, geophysical and geopolitical knowledge. Examples include coastlines and political boundaries, mining activity information, crustal depth, locations of sensor network components, and the average speed of sound in the ocean for a given location. Contextual data commonly has a spatial component which ties attribute information of a *data set* feature to a known location on the surface of the Earth. Spatial relationships may be found to exist between events and contextual information, which can facilitate the evaluation of an event.

For *Knowledge Base* purposes, these data are generally grouped and organized using an application called ArcView GIS, from Environmental Systems Research Institute, Inc., (ESRI). It is the standard geographic information system software selected for use in the GNEM R&E program. Contextual *data sets* submitted to the *Knowledge Base* should be delivered in a format suitable for use by this software. Following is a list of acceptable formats for use with the ArcView GIS application. Some may require specific extensions to the software, which are either free or may require the purchase of a separate user license. If necessary, *Contributors* are encouraged to contact the *Product Integrator* to discuss alternatives.

If a *Contributor* is providing imagery data, the following standard formats can be used:

ARC Digitized Raster Graphics (ADRG) (requires ArcView's ADRG Image Support extension)
BMP
BSQ, BIL and BIP
Compressed ARC Digitized Raster Graphics (CADRG) (requires ArcView's CADRG Image Support extension)
Controlled Image Base (CIB) (requires ArcView's CIB Image Support extension)
ERDAS
GRID (ESRI proprietary data format)
IMAGINE (requires ArcView's IMAGINE image extension)
IMPELL Bitmaps (Run-length compressed files)
Image catalogs
JPEG (requires ArcView's JPEG image extension
MrSID (requires ArcView's MrSID image extension)
National Image Transfer Format (NITF) (requires ArcView's NITF Image Support extension)
Sun rasterfiles
TIFF/GeoTIFF
TIFF/LZW compressed (requires optional software library for viewing)

ArcView GIS supports the use of vector and raster data in any of the following formats:

ArcView shapefiles (.shp)
ARC/INFO coverages
ARC/INFO interchange files (.e00)
ARC/INFO grids
USGS Digital Elevation Model (DEM) raster files
ASCII raster files
Binary raster files
CAD drawings
ArcSDE data
TINs (Requires purchase of the 3D Analyst extension)
Vector Product Format data (NIMA data format)

ArcView also supports access to tabular information:

ARC/INFO INFO tables
Oracle (or other database server) tables (export format or flat files)
dBASE III and dBASE IV files
Comma- or tab-separated text files

3.4 Supporting Information

Supporting information are models, algorithms or papers that are available for reference event data, parametric grid data or contextual data. Some examples of these kinds of supporting information are travel time tables, velocity models, dispersion tables, attenuation tables, parameter files, discrimination routines, magnitude formulas, attenuation formulas, meteorological models and propagation models.

Since the formats for supplemental information can be variable, the *Contributor* needs to coordinate with their *Product Integrator* on the best format to use. In general, files written in Word, Word Perfect or Framemaker, plus ASCII text files are acceptable formats.

3.5 Research Tools

Research Tools are used to help create and access *research products* in the *Knowledge Base*. They force consistency in *research products* from a variety of *Contributors*, and promote better testing of the *research products* by the *Product Integrators* in their *Scientific Integrator* role, *KB Integrators* and the KB customer.

A research tool can be written in the language of the *Contributor's* choice. *Contributors* are encouraged to develop any research tools in an environment similar to the *Product Integrator's*. All the files needed to compile the program need to be packaged together, along with the instructions for compiling the program and a user's manual for the tool. In addition, the *Product Integrator* will need to know the environment under which the tool was developed. This includes the required hardware and required software: OS, OpenGL, compilers, etc.

4.0 Documentation and Metadata

Every *Contributor* of a *research product* will need to provide documentation on their product. This documentation will be used by the *Product Integrator* to write the *research product* documentation and *metadata* necessary if the *research product* is included in the *Knowledge Base*.

Metadata answers the questions who, what, when, where, why and how about every facet of the *research product* that is being documented. It describes the origins of the *research product* and tracks any changes that are made to it. For example, the legend of a map is considered pure *metadata*. The legend describes **who** published the map, **when** it was published, **what** kind of map it is and **what** it describes, **where** the information is (i.e. spatial reference), and **how** the map was compiled (scale and accuracy). The same kind of descriptive information is needed for all *research products*. The *metadata* for the *Knowledge Base* is based on the "Content Standards for Digital Geospatial Metadata", the standard developed by the Federal Geographic Data Committee (FGDC) in 1998, which can be found at http://www.fgdc.gov/metadata/contstan.html. The FGDC document is the standard for *metadata* that Federal agencies are instructed to use to document new geospatial data. A committee at Sandia National Laboratories (SNL) went through the FGDC standard and chose the pieces of *metadata* necessary to adequately describe the *data sets* expected in the *Knowledge Base*.

For a *Product Integrator* to be able to write both the necessary documentation and enter the *metadata* into the *Knowledge Base* Catalog, a *Contributor* must answer the following questions, and give an electronic copy to the *Product Integrator*. Please consult with your *Product Integrator* on the level of useful detail for each product summary (e.g. one for each map or one for the overall *research product*). You will need to use one of the following word processors: Word, Word Perfect or Framemaker. Templates can be found in the Appendix.

1. Information on the *Contributor* - name, organization, address, phone, fax and email
2. Name and Version of the *research product*
3. If there is an associated document or report, provide its title in full bibliographic citation form
4. Provide the research contract number and name of the sponsoring organization
5. Provide a short paragraph with a high level description of the *research product*.
6. List any guidelines or warnings about the *research product* the user needs to be aware of (caveats).
7. List any information the user needs to be aware of to use the *research product* in a proper manner (dependencies).
8. If appropriate for the *research product*, list one or more geographic names (e.g. North America, Mexico) that describe the location covered by the *research product*.
9. List one or more common words or phrases that can be used to describe the subject of the *research product*.
10. If appropriate for the *research product*, list the geographic bounding coordinates (upper and lower bounds)
11. If appropriate for the *research product*, list the latitude and longitudes resolution.
12. If appropriate for the *research product*, indicate the vertical frame of reference (depth or elevation), the units used and the resolution.

13. For a *research product* that is event data and/or parametric grid data, indicate the time frame in which the events occurred. For example, if the *research product* is a set of amplitude measurements on events that occurred between January 1, 1995 and October 15, 1997, please list the time frame as start date = 1/1/95 and end date = 10/15/97.
14. For a *research product* that is contextual data, supplemental information or research tools, provide the publication/release date.
15. List any people and/or organizations that contributed to the *research product*.
16. Describe the processing environment within which the *research product* was produced. This should include the hardware, the version of the operating system and versions of software.
17. Define the system of objects used to represent space in the *research product* as point, vector or raster.
18. List all the sources (title, originator, date, publication, if any) used in developing this *research product*. These sources should include the raw data, computer source code and methodology papers. Consider this to be the "Data" portion of a "Data and Methods" section in a paper.
19. Describe the steps that were taken to go from the raw data to the finished product. Consider this to be the "Methods" portion of a "Data and Methods" section in a paper.
20. Discuss the tests and results obtained for determining the accuracy of the values in the *research product*.
21. Provide information about omissions, selection criteria, generalizations, definitions used and any other rules used to derive the *research product*.
22. List all the electronic files that make up the *research product*.

In addition to answering the questions, if there is a report on the *research product*, the *Contributor* should also provide an electronic (if possible) copy to the *Product Integrator.*

5.0 Summary

The Knowledge Base Contributor's Guide is intended to facilitate getting high quality input from the *Contributor* to the *Product Integrator*. Key points that a reader should get by reading this guide are:

1. The *Product Integrator* is the point of contact for a *Contributor*. (Section 2.0)

2. There are five broad categories of *data sets* that are included in the *Knowledge Base*: event data, parametric grid data, contextual data, supporting information and research tools. A *research product* consists of one or more *data sets*. There are standard formats for the different categories of *data sets*. (Section 3.0)

3. In order for the *Product Integrator* to complete the required documentation and *metadata* for a *research product*, there are specific questions about the *research product* and how it was produced that must be answered by the *Contributor*. The *Contributor* can use one of the templates provided in this document to provide the answers to the questions. (Section 4.0, Appendix).

6.0 References

Federal Geographic Data Committee, 1998. **Content Standard for Digital Geospatial Metadata (revised June 1998)**, Washington, D.C., FGDC-STD-001-1998. (also available from http://www.fgdc.gov/metadata/contstan.html)

Gallegos, D. P., D. B. Carr, P. B. Herrington, J. M. Harris, C. L. Edwards, S. R. Taylor, J. J. Zucca, D. B. Harris, N. A. Wogman, D. N. Anderson and L. A. Casey, 2002. **The Integration Process for Incorporating Nuclear Explosion Monitoring Research Results into the National Nuclear Security Administration Knowledge Base**, Sandia Report SAND2002-2772. (also available from http://www.nemre.nn.doe.gov/nemre/kbase_info.html)

7.0 Glossary of Terms

The *Knowledge Base* is among other things, an internally consistent construct for which we have chosen to use specialized definitions of words or phrases (highlighted with italics). This glossary provides definitions for these, and it should be referred to as needed by the reader when an italicized term is encountered in the guide. A more extensive glossary of terms associated with integration of *research products* into the *Knowledge Base* can be found in Gallegos et al. (2002)

Contributor

An individual or a group of individuals creating *data sets* and incorporating them into *research products* for delivery to a *Product Integrator*. A *Contributor* may be from a NNSA or DOE laboratory, a university or from the private sector.

data set

A *data set* is a collection of data usually all of the same kind.The five broad categories of *data sets* are event data, parametric grid data, contextual data, supporting information and research tools. A collection of *data sets* frequently makes up a *research product*.

Knowledge Base

An NNSA-sponsored integrated, organized collection of *research products* and Information Products used for automated regionalization and analyst efforts delivered to the USNDC.

KB Coordinator

An individual or group responsible for the oversight and integration of *data sets* and *research products* into Information Product deliveries for potential KB population. The *KB Coordinator* works closely with and provides Information Products to the *KB Integrator*. For the definition of the *KB Coordinator's* duties, see Gallegos et al. (2002).

KB Integrator

Individuals who combine the Information Products into the KB product. SNL acts as the *KB Integrator*. For the definition of the *KB Integrator's* duties, see Gallegos et al. (2002).

metadata

Information that describe the content, quality, condition and other characteristics of data, especially geospatial data. The *metadata* for the *Knowledge Base* is based on the "Content Standard for Digital Geospatial Metadata", http://www.fgdc.gov/metadata/contstan.html.

Product Integrator

A scientist assigned by the NNSA to be responsible for coordinating with specific research and development contractors, particularly those not funded by NNSA. The *Product Integra-*

tor's responsibility is to determine whether the contract's results and products could or should be considered for incorporation into the *Knowledge Base* and to facilitate transfer of the information through a *KB Coordinator*. For the definition of the *Product Integrator*'s duties, see Gallegos et al. (2002).

research product

A product of contracted research. A *research product* frequently consists of one or more *data sets* and associated software. Each *research product* must have corresponding *metadata*. Typically, several *research products* go into a single Information Product.

Appendix - Contributor's Research Product Summary for Product Integrators

For a *Product Integrator* to be able to write both the necessary documentation and enter the *meta-data* into the *Knowledge Base* Catalog, a *Contributor* needs to answer the following questions, and give an electronic copy to their *Product Integrator*.There are three different templates that can be used: Microsoft Word, Word Perfect and Framemaker.

1. Provide the following information:

Contributor's Name:
Organization:
Mailing Address (including mail code if applicable):
Phone:
Fax:
Email:

2. Provide the name and version of the research product:

3. If there is also an associated documentation report, provide its title in full bibliographic citation form.

4. Provide the research contract number and name the sponsoring organization for the research product.

5. Provide a short paragraph with a high level description of the research product.

6. Please list any guidelines or warnings about the research product the user needs to be aware of (i.e. caveats):

7. Please list any information the user needs to be aware of to use the research product in a proper manner (i.e. dependencies):

8. If appropriate for the research product, list one or more geographic names (e.g. North America, Mexico) that describe the location addressed by the research product:

9. List one or more common words or phrases that can be used to describe the subject of the research product:

10. If appropriate for the research product, list the geographic bounding coordinates. **(0 to +90 for northern latitude, 0 to -90 for southern latitudes; 0 to +180 for eastern longitudes, 0 to -180 for western longitudes)**:

Upper Bound Latitude Coordinate:
Lower Bound Latitude Coordinate:
Eastern Most Longitude Coordinate:
Western Most Longitude Coordinate:

11. If appropriate for the research product, list the latitude and longitudes resolution:

Latitude resolution:
Longitude resolution:

12. If appropriate for the research product, indicate the vertical frame of reference (depth or elevation), the units used and the resolution:

Frame of reference:
Units used (km, m, etc.):
Resolution:

13. For a research product that is event data and/or parametric grid data, indicate the time frame in which the events occurred. For example, if the research product is a set of amplitude measurements on events that occurred between January 1, 1995 and October 15, 1997, please list the time frame as start date = 1/1/95 and end date = 10/15/97.

Start date:
End date:

14. For a research product that is contextual data, supplemental information or research tools, provide the publication/release date:

Publication date:

15. List any people and/or organizations that contributed to the research product:

16. Describe the processing environment within which the research product was produced. This should include the hardware, the version of the operating system and versions of software.

17. Define the system of objects used to represent space in the research product as point, vector or raster:

18. List all the sources (title, originator, date, publication, if any) used in developing this research product. These sources should include the raw data, computer source code and methodology papers. Consider this to be the "Data" portion of a "Data and Methods" section in a paper.

19. Describe the steps that were taken to go from the raw data to the finished product. Consider this to be the "Methods" portion of a "Data and Methods" section in a paper.

20. Discuss the tests and results obtained for determining the accuracy of the values in the research product:

21. Provide information about omissions, selection criteria, generalizations, definitions used and any other rules used to derive the research product:

22. Please list all the electronic files that make up the research product.

Distribution

Sandia National Laboratories Internal:

110	MS1138	Dorthe Carr, 6533
1	MS9018	Central Tech Files, 8945-1
2	MS0899	Technical Library, 9616
1	MS0612	Review & Approval Desk, 9612